Discovering Hidden Pathways in Bioinformatics

Paulo J.G. Lisboa[1,⋆], Ian H. Jarman[1], Terence A. Etchells[1], Simon J. Chambers[1], Davide Bacciu[2], Joe Whittaker[3], Jon M. Garibaldi[4], Sandra Ortega-Martorell[5], Alfredo Vellido[6], and Ian O. Ellis[7]

[1] School of Computing & Mathematical Sciences, Liverpool John Moores University, UK
[2] Dept. of Computer Science, University of Pisa, Italy
[3] Dept. of Mathematics and Statistics, University of Lancaster, UK
[4] School of Computer Science, University of Nottingham, UK
[5] Dept. of Biochemistry and Molecular Biology, Universitat Autònoma de Barcelona, Spain
[6] Dept. of Computer Languages and Systems, Universitat Politècnica de Catalunya, Spain
[7] Dept. of Histopathology, School of Molecular Medical Sciences Nottingham University Hospitals Trust, University of Nottingham

Abstract. The elucidation of biological networks regulating the metabolic basis of disease is critical for understanding disease progression and in identifying therapeutic targets. In molecular biology, this process often starts by clustering expression profiles which are candidates for disease phenotypes. However, each cluster may comprise several overlapping processes that are active in the cluster. This paper outlines empirical results using methods for blind source separation to map the pathways of biomarkers driving independent, hidden processes that underpin the clusters. The method is applied to a protein expression data set measured in tissue from breast cancer patients (n=1,076).

Keywords: clustering, independent components, hidden sources.

1 Introduction

Disease sub-typing is a priority for interventional medicine. A commonly used first step to find sub-types in bioinformatics is to cluster expression profiles, searching for disease phenotypes by grouping observation into naturally recurring patterns. An important aim of disease phenotyping is to gain insights into the mechanisms that drive metabolic function. These are commonly represented

⋆ The authors acknowledge the support of Dr A. Green and other members of the Breast Cancer Research Team of the University of Nottingham, UK, and Dr. G. Ball, of Nottingham Trent University, UK, towards the collection of high-throughput protein expression dataset used in this study, and financial support from the European Network of Excellence Biopattern (FP6-2002-IST-1 No. 508803). A. Vellido is supported by Spanish R+D project TIN2009-13895-C02-01.

E. Biganzoli et al. (Eds.): CIBB 2011, LNBI 7548, pp. 49–60, 2012.

as conditional independence networks that map the multivariate associations between the expression levels of molecular biomarkers.

We hypothesise that individual clusters involve different biological pathways, independently active, which combine in different proportions to form each cluster profile. In mathematical terms, we seek independent components whose mixing coefficients separate when labelled by cluster membership.

Independent Component Analysis (ICA) has been applied to microarray data as a method of unsupervised analysis, reporting a significant improvement in finding biologically relevant transcriptional models [1,2]. Furthermore, expression modes and so-called meta-modes were also derived from microarray data by Lutter and colleagues [3], who remarked that deep exploration by application of ICA is still needed but it is an appropriate tool to uncover underlying biological mechanisms from molecular data. This is mirrored in more recent work by Schwartz and Shackney [4].

However, in these studies each of the identified sources is described by a fixed expression profile, in effect a row of covariates with the same dimensionality as the data. We propose a procedure to identify the conditional independence map for each source. The ultimate aim of this analysis is to elucidate the different biological processes that underpin biological function and explain the contribution of each process to the expression profile of each phenotype.

2 Materials and Methods

2.1 Data

We used a previously studied dataset (n=1,076) with 25 protein expression values. The measurements were made from tissue samples collected at initial excision surgery for breast cancer [5]. They are listed in Table 2.1.

2.2 Existing Methods

Clustering methods commonly used in bioinformatics include k-means, Partition Around Medoids and hierarchical algorithms. All of these may be used alone or in combination to form consensus clusters [6]. A fundamental question is whether it is possible to systematically generate cluster solutions that are reproducible in the sense that repeating the analysis will produce approximately the same solutions. A possible approach is to choose a method with a unique outcome, for instance hierarchical clustering. However, with agglomerative methods early stage errors can arise, leading to sub-optimal solutions as compared with partition algorithms [7]. This is especially the case with high dimensional data due to ultrametricity. A robust methodology was applied to the k-means algorithm, by mapping the landscape of solutions obtained for different initializations, using twin directions of within-cluster separation and between-cluster stability [8], to choose a reproducible solution that scores highly in both performance indices.

Low-dimensional visualization of high dimensional data is achieved effectively using the axes defined by the eigenvalues of the separation matrix consisting